Progressively Spiral Mamba Fusion for Multimodal Tracking

Zixuan Wang[1], Baojie Fan[1]*, Jiajun Ai[1], and Wenzhang Zhou[1]

Nanjing University of Posts and Telecommunications, Nanjing, China
jobfbj@gmail.com

Abstract. Multimodal object tracking has received extensive attention due to its strong robustness and complementary collaboration. However, most existing methods only evaluate the cross-modal fusion at specific multi-modal layers, and overlook the interaction among different feature layers from single modal, which limits the utilization and propagation of cross-modal complementary information and consequently degrades tracking performance. To alleviate this issue, we propose a progressive multimodal tracker based on State Space Model (SSM) named PSMTrack, which consists of three key modules. First, Difference-guided Bidirectional State-space Mamba Enhancement (DBME) is developed to leverage difference-aware cues and bidirectional SSM modeling to progressively enhance modality representations for subsequent tracking, while suppressing background and similar distractions. Second, we adopt an expert routing mechanism to adaptive select a subset of layers and perform cross-layer interaction to alleviate the information loss caused by relying solely on the last layer. Finally, we design a Hybrid Spiral Mamba Fusion (HSMF) module that performs dynamic spatial modeling with spiral Mamba from both local and global scales for final cross-modal fusion, thereby capturing long-range dependencies with near-linear complexity and improving bounding-box regression stability. Extensive experiments on multiple mainstream RGB-X benchmarks validate that our model consistently improves both accuracy and robustness.

Keywords: Multimodal tracking · Vision Transformer · Mamba

1 Introduction

The purpose of Single Object Tracking (SOT) [34, 41] is to continuously localize the position of the target throughout a video sequence, based on the initial property of the tracked target provided in the first frame. It has been widely applied in robot operation [31], autonomous driving [44], and intelligent supervision [45]. Although many RGB-based trackers [2, 5, 23] have achieved robust astonishing performance and significant improvement in generalization ability. However, when encountering extreme conditions such as strong occlusion, extreme lighting

* Corresponding author.

conditions, fast motion with deformation, and background distractions [4,43,47], RGB-based trackers are prone to losing the targets.

To overcome above challenges and enhance robustness, a natural choice is to introduce the complementary of multi-source sensing. Thermal infrared (TIR) [16,19] can produce stable imaging under low-light conditions, depth (D) [37,49] can provide a hierarchical structure and geometric cues under occlusion, event cameras (E) [51] maintain high temporal resolution even during fast motion. The fusion of RGB with above modalities (collectively referred to as RGB-X) can leverage the complementary strengths of their respective modalities in extreme tracking scenarios, enabling trackers to achieve satisfactory accuracy and robustness. Additionally, more and more multimodal datasets and evaluation benchmarks have further propelled the development of multimodal trackers.

Specifically, existing multimodal tracking methods use cross-modal attention or adapters to directly concatenate or weight features from different modalities into a unified representation [1,11,25]. Although these methods can achieve the fusion between modalities, they still have some limitations. First, to reduce the computational and memory overhead, cross-modal attention interaction is often only carried out in partial layers, resulting in insufficient utilization of inter-modal information. Second, due to adapters are parameter-light, the interaction capability of this type of method is relatively weak. To achieve sufficient interaction effects often requires repeated deployment over many layers and long token sequences, which forces a trade-off between the number of inserted layers and the sequence length. Meanwhile, existing methods ignore the cross-layer complementarity: high-level features provide stable semantics, whereas low-level features provide detailed texture details. Notably, recently emerging State Space Models (SSMs) [9, 26], particularly Mamba [8], as a new sequence model, has been introduced into many object tracking models [12, 22]and has demonstrated highly competitive performance.

Inspired by the above, we propose a progressive multimodal tracker based on the State space models (SSMs), comprising three innovative modules: Difference-guided Bidirectional State-space Mamba Enhancement (DBME), Cross-Layer Interaction (CLI) and Hybrid Spiral Mamba Fusion (HSMF). Specifically, after each Transformer block, differential features are constructed based on two modalities, model them with Mamba, and inject back into the current layer in a residual manner. This layer-by-layer refinement promotes the effective transmission of complementary information. Meanwhile, we construct a cross-layer interaction module that dynamically selects and aggregates multi-level features through a lightweight sparse routing mechanism. This module simultaneously considers high-level semantic information and low-level texture details, and compensates for the loss of temporal information caused by relying solely on the final layer. In the final fusion stage, the Hybrid Spiral Mamba Fusion (HSMF) module is introduced to conduct dynamic spatial modeling of the fused features across local and global scales, achieving better boundary details and global consistency with near-linear computational complexity. Through extensive experiments, we validate the effectiveness of our model.

Our main contributions can be summarized as follows:

• We propose a progressive multimodal tracker based on Spiral Mamba that enables efficient cross-modal fusion at both local and global scales while maintaining linear computational complexity.

• We design the DBME module to fully utilize the differential information of multiple modalities during the feature extraction stage, and propose the CLI module to adaptively aggregate multi-layer features and mitigate the information loss caused by relying solely on the final layer.

• Through extensive experiments, our model achieves new state-of-the-art (SOTA) performance on five mainstream RGB-D/T/E tracking benchmarks.

2 Related Work

2.1 Multimodal Tracking

Multimodal object tracking [12,42] aims to continuously locate targets throughout video sequences by jointly leveraging RGB and one or more auxiliary modalities such as depth, thermal infrared, and event date [17,33,38], and to improve the robustness of the tracking network under severe target appearance variations. Compared with single-modality tracking that relies only on appearance cues, multimodal tracking can provide complementary information in extreme scenarios such as occlusion, low illumination, fast motion, and camera scale changes. Deep helps reduce uncertainty in scale and distance, thermal radiation against light degradation, and event data captures motion edges with high temporal resolution.

In recent years, several work has begun to explore a unified architecture capable of handling a variety of multimodal tracking tasks. For example, ViPT [46] proposed the method of prompt learning, which no longer treats the X modality as an equally weighted parallel branch but as a conditional prompt injection into the RGB backbone. SDSTrack [11] adopt parameter-efficient adapters to realize symmetrical interaction and fusion among modalities at multiple stages, and transfer complementary cues across different layers. OneTracker [10] proposes a universal framework for unified visual object tracking. By pretraining the Foundation Tracker on large-scale RGB data and then adapting different modalities through prompt tuning, t achieves efficient unified modeling. STTrack [12] further introduces the temporal state generator of cross-modal Mamba, which continuously produces temporal-information tokens to unify the modeling of inter-frame history and cross-modal signals.

2.2 State Space Model

State Space Model (SSM) [9], as a method for sequence modeling, has received widespread attention in recent years due to its ability to capture long-range dependencies while maintaining linear computational complexity. Building on this, Mamba [8] introduces a selective mechanism and has been adopted in many computer vision tasks.

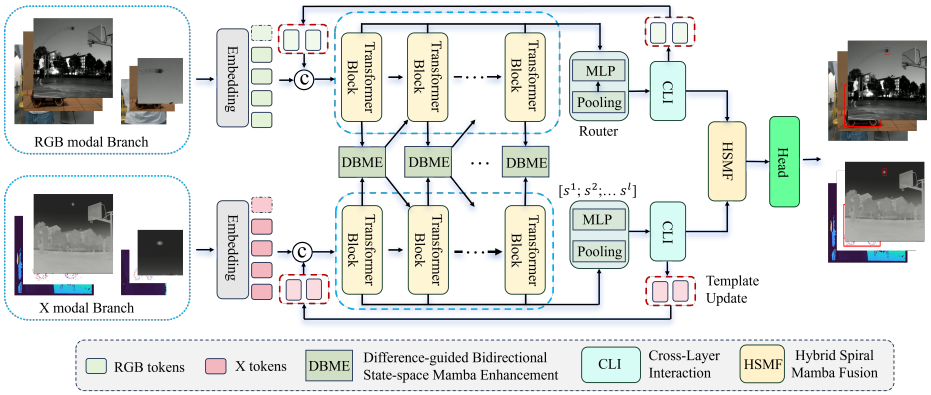

Fig. 1: The overall structure of our PSMTracker. Firstly, RGB and X images are converted into token sequences and fed into Transformer blocks. After each Transformer block, the RGB and X features are processed by the DBME module to enhance their representations and then passed to the next block. Then, modality-specific routers select subsets of layer features and feed them into the CLI module for cross-layer interaction. Meanwhile, a template update mechanism is employed to maintain and refine the historical target representation during tracking. Finally, the two modal features are fused by the HSMF module and fed into the tracking head for target localization.

Vim [48] designs a bidirectional state space with positional embeddings to better model long sequences. VideoMamba [21] introduces Mamba into video domain to address local redundancy and global dependency issues. In the visual tracking task, MambaLCT [22] scans search frames along the time with Mamba, selectively compressing the information from historical frames into the hidden state to build a long-term context from the first frame to the current frame. MCITrack [13] takes the hidden state of Mamba as a video-level information channel and deeply injects cross-frame context into the feature extraction process, conveying richer temporal information with lower additional token overhead. MamTrack [30] introduces an efficient modal selection fusion module based on Mamba for RGB-Event tracking. Coupled with a history decoder that models target appearance changes over time, it unifies multimodal spatial features with long-range temporal cues.

More recently, VSSD [28] further eliminated the causal mask in the state-space duality (SSD) of Mamba2 [7], introducing a non-causal state-space duality (NC-SSD) that further improves global context modeling for vision tasks. We further improve NC-SSD by proposing a cross-layer interaction strategy that aggregates multi-level features for visual tracking scenarios.

3 Method

In this paper, we propose a novel multimodal visual object tracking method named PSMTrack, which utilizes Mamba to achieve efficient interaction and

fusion of multimodal features. As shown in Fig. 1, the overall architecture of PSMTrack is presented. In this section, we first briefly review state space model. Subsequently, we provide a detailed introduction to each module of PSMTrack.

3.1 Preliminaries

State Space Models (SSMs) [9] restores sequence modeling to a process of "input - dynamic system - output".The system is prototypical of continuous linear time-invariant (LTI) dynamics, absorbing historical information via a latent state and generating current representation. The typical form is:

$$\dot{h}(t) = A\,h(t) + B\,x(t), \quad y(t) = C\,h(t), \tag{1}$$

where $\dot{h}(t)$ denotes the time derivative of $h(t)$, $A \in \mathbb{R}^{N\times N}$ is the dynamics parameter, and $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{1\times N}$ are the projection parameters.

When dealing with discrete sequences such as text or images, in order to match the discrete nature of the data, for continuous-time signals, the state space models uses zero-order hold (ZOH) to convert them into discrete-time signals. Let Δ be a predefined time-scale parameter used to convert the continuous-time parameters A, B into their discrete-time counterparts $\bar{A}, \bar{B}$, formally defined as:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}\big(\exp(\Delta A) - I\big)\,\Delta B. \tag{2}$$

Mamba [8] breaks the LTI limitation of SSM by introducing the a selective state space mechanism. It makes the model parameters dependent on the input data, thereby enhancing the model's contextual awareness and improving the ability to handle long sequences.

3.2 Difference-guided Bidirectional State-space Mamba Enhancement

In multimodal target tracking, since the RGB and X modals originate from different imaging mechanisms, their differential features usually contain rich complementary clues. However, many current works directly interact with features of different modalities through the transformer, which has a high quadratic computational complexity. At the same time,many existing works also make insufficient explicit use of the difference information between modalities. To solve this problem, we propose a Difference-guided Bidirectional State-space Mamba Enhancement (DBME) module, as depicted in Fig. 2, deploy it after each Transformer block, operated layer by layer in 12 layers, simultaneously enhancing the self-expression and complementary injection capabilities of the two modals without changing the shape of the sequence.

Specifically, inspired by [48], bidirectional scanning is more capable of generating stable and consistent responses at spatial boundaries than unidirectional scanning. Therefore, we respectively conduct forward SSM and reverse SSM scans on the visual tokens of the two modalities to achieve bidirectional modeling of the original features and enhance their own feature expression.

In parallel, for the two modal features from the same layer, we perform a absolute subtraction to obtain the difference feature of the same layer F_i^d, which explicitly represents the inconsistent regions and complementary cues between the two modalities. Then, the difference feature F_i^d is selectively modeled as a sequence by a Mamba block, and a tanh activation is applied to suppress noise from the common mode, highlighting the truly informative cues that need to be injected. Finally, the gated bidirectional response is injected back into the original features in a residual manner, ensuring stable and controllable training while achieving progressive layer-wise feature enhancement.

We denote the two modality features at layer i as F_i^r and F_i^x, the processing of this module can be expressed as follows:

$$F_i^d = |F_i^r - F_i^x|\,, \tag{3}$$

$$F_i^{r\text{-}f} = \Phi_{\mathrm{SSM}}(F_i^r)\,, \quad F_i^{r\text{-}b} = \mathrm{flip}(\Phi_{\mathrm{SSM}}(\mathrm{flip}(F_i^r)))\,, \tag{4}$$

$$F_i^{x\text{-}f} = \Phi_{\mathrm{SSM}}(F_i^x)\,, \quad F_i^{x\text{-}b} = \mathrm{flip}(\Phi_{\mathrm{SSM}}(\mathrm{flip}(F_i^x)))\,, \tag{5}$$

$$F_i'^r = \Phi_L\Big(\Big(F_i^{r\text{-}f} + F_i^{r\text{-}b}\Big) \cdot \tau\big(\mathrm{Mamba}\big(F_i^d\big)\big)\Big) + F_i^r\,, \tag{6}$$

$$F_i'^x = \Phi_L\Big(\Big(F_i^{x\text{-}f} + F_i^{x\text{-}b}\Big) \cdot \tau\big(\mathrm{Mamba}\big(F_i^d\big)\big)\Big) + F_i^x\,. \tag{7}$$

where $\Phi_L(\cdot)$ and $\tau(\cdot)$ denote the linear projection and the tanh activation function, respectively. flip$(\cdot)$ denotes reversing the order of a sequence.

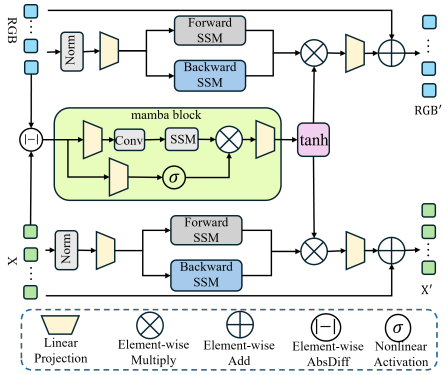

Fig. 2: The detailed structure of the DBME module. The feature discrepancy between the two modalities is first modeled with Mamba, and the result is then added back to the original features after bidirectional state space modeling through the residual method to obtain the enhanced features of each branch respectively.

3.3 Cross-Layer Interaction

Existing feature aggregation methods suffer from two critical limitations. First, relying solely on the final layer output leads to information forgetting. Second,

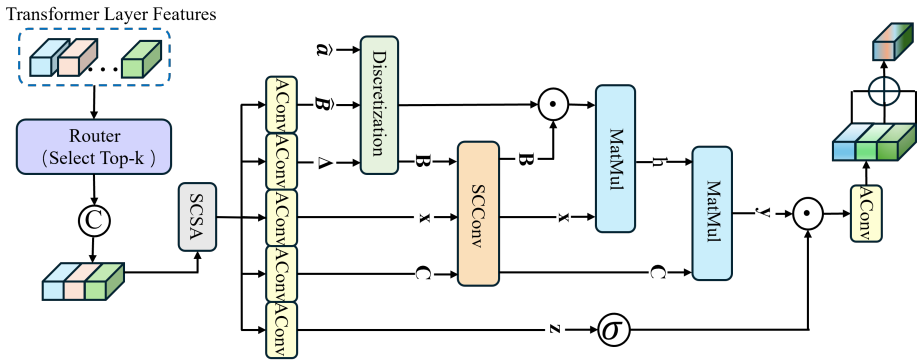

Fig. 3: Architecture of CLI. Cross-layer interaction across multiple feature levels is achieved through a variant of NC-SSD.

fixed layer selection strategies fail to adapt to dynamic scene variations. To address these issues, we propose a Cross-Layer Interaction (CLI) module Fig. 3 that dynamically aggregates features from multiple layers of each modality, enabling more robust feature fusion.

We first adopt a routing-based sparse selection strategy to dynamically select a subset of layer-wise features. A modality-shared router R aggregates representations from all Transformer layers to produce layer selection scores. The final-layer feature is always retained as a shared expert:

$$s^n = R\left([F_1; F_2; \ldots; F_L]\right), \tag{8}$$

$$R(x) = \mathrm{MLP}\left(P(x)\right). \tag{9}$$

where $P(\cdot)$ denotes the pooling operation, and MLP refers to a multilayer perceptron.

The selected features are then concatenated along the channel dimension and fed into the CLI module. Since direct cross-layer concatenation may introduce inter-layer misalignment and redundant activations, we apply SCSA [29] to the concatenated features along both spatial and channel paths for explicit denoising and recalibration, thereby reducing the sensitivity of the subsequent dynamic generator to noise. Compared with the linear layers in the original NC-SSD, an adaptive convolution (AConv) 1×1 will dynamically adjust the weights according to the input features for adaptive adjustment, enabling the model to adaptively fuse features from different layers. This process allows the model to focus on the most relevant features, improving SSD's spatial modeling capability when dealing with cross-layer features.

For B, C, and x, we use SCConv [20] instead of DWConv for processing. Under the premise of maintaining the shape of the 3×3 local sensory field and tensor unchanged, a more stable spatial reconstruction effect can be achieved than depthwise convolution, and the bias caused by spurious activation can be reduced under occlusion, blur, and weak-texture backgrounds. Subsequently, we

perform discretization and then conduct modeling through multiple Hadamard (element-wise) products.

Finally, a linear projection of AConv1×1 is performed, followed by a nonlinear activation, to obtain the per-layer interaction features $F_{\text{interaction}}$. The $F_{\text{interaction}}$ terms are then combined by equal-weighted summation across layers to produce the final cross-layer fused feature:

$$F = \sum_{l \in S} F^{l}_{\text{interaction}}, \tag{10}$$

where S denotes the set of selected layers determined by the router. After completing the interaction, each modality's F serves as the input feature for the fusion between the modalities.

3.4 Hybrid Spiral Mamba Fusion

After completing the layer-by-layer differential enhancement and cross-layer interaction at the end, since the RGB modality and X modality (such as thermal infrared/depth/event) often focus on different parts of the scene in space, relying on a single modality alone may still lead to semantic fragmentation and local instability. To this end, we introduce the Hybrid Spiral Mamba Fusion (HSMF) module before the final prediction. This module is built around the State Space Model (SSM) with linear computational complexity, through sequential modeling by cascading the local and global spiral Mamba, ensure both the reliability of local details and global consistency, and gradually compress the multimodal features into a unified and robust feature representation.

As shown in Fig. 4, given the RGB modality feature F^{RGB} and the auxiliary X modality feature F^{X}, We first concatenate them along the channel dimension at the same spatial coordinates to obtain the initial multimodal feature F:

$$F = \text{Concat}\left(F^{\text{RGB}}, F^{\text{X}}\right). \tag{11}$$

The RGB and X modality responses at the same position are merged into the same high-dimensional feature vector, all subsequent processing is carried out on this fusion feature. In order to simultaneously obtain fine-grained boundary reliability and globally consistent semantics without relying on global self-attention (whose computational complexity increases quadratically with spatial scale), we process the feature F using Mamba in two sequential stages. The local Mamba stage performs noise suppression and boundary refinement within a small range, while the global Mamba phase establishes long-term dependencies and identity consistency across the entire feature map. The final fused feature output after the HSMF is denoted as F^{fusion}, and is directly used as the input of the regression head to predict the target bounding box.

Local Mamba. The purpose of the local Mamba modeling stage is to enhance the local consistency of the feature F, that is, to determine within a limited spatial neighborhood which responses are stable and reliable real targets, and which are noise generated by a single modality. Specifically, we first divide the

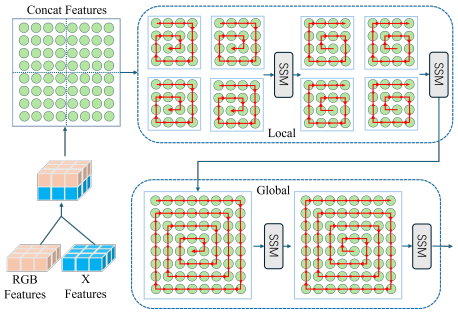

Fig. 4: Illustration of HSMF. The module first conducts local feature interaction via the local spiral Mamba, and then the global spiral Mamba further enables global feature interaction.

feature F uniformly into non-overlapping regions of size $k \times k$. For each region, we process it in a spiral scanning order. The features are progressively accumulate along the spiral path and are selectively preserved or suppressed, this enables the modality discrepancies to be aligned automatically within each $k \times k$ region, only the boundaries and textures supported by both modalities are preserved and amplified, while noise or false boundaries caused by a single modality are suppressed. In order to achieve comprehensive information exchange, we perform fusion along the spiral directions from the inside to outside and from the outside to inside simultaneously.

After the local Mamba stage, the features in each local spatial region have been changed from a simple sum of modalities to clean and salient responses after modalities mutually correction.

Global Mamba. The global Mamba modeling stage takes as input the features produced by the local Mamba stage, and builds consistency and global contextual modeling for the entire feature map. Make sure that the network can not only clearly see in a local area whether a region belongs to the target, but also maintain the judgment that this is the same target on a global scale, rather than similar distractions in the scene. Unlike the local stage, we no longer divide the feature map;instead, we treat the entire feature map as an integrated region, and process the entire $H \times W$ feature map using a spiral scanning scheme.

This spiral sequence ensures that adjacent positions in the sequence are also relatively close in the two-dimensional plane, thereby allowing the state space model to undergo smooth, directionally continuous long-range propagation over the entire feature map, without breaking spatial coherence as would occur with a randomly shuffled sequence. In this stage, regions affected by partial occlusion or severe illumination changes can receive consistent target descriptions from unaffected regions, avoid the loss of identity consistency caused by local corruption. Finally, we reshape the sequential output back to obtain the globally consistent fusion feature F^{fusion}, which is then directly fed into the prediction head to complete the prediction of the target box.

3.5 Head and Loss

We follow most of the tracking methods [35, 40], use and freeze a classic fully convolutional network (FCNs) prediction head. We use L_{cls} [15] to denote the weighted focal loss for the classification task, The generalized IoU loss L_{iou} [27] and the L1 loss are used for bounding box regression. The overall loss function is formulated as follows:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_1 L_{\text{iou}} + \lambda_2 L_1, \tag{12}$$

where $\lambda_1 = 2$ and $\lambda_2 = 5$ are the regularization coefficients that balance the contributions of each loss term.

4 Experiment

Method	LasHeR [19]		RGBT234 [17]	
	PR ↑	SR ↑	MPR ↑	MSR ↑
PSMTrack(Ours)	**76.5**	**61.2**	**91.2**	**67.3**
STTrack [12]	76.0	60.3	89.8	66.7
BAT [1]	70.2	56.3	86.8	64.1
OneTracker [10]	67.2	53.8	85.7	64.2
SDSTrack [11]	66.5	53.1	84.8	62.5
ViPT [46]	65.1	52.5	83.5	61.7
ProTrack [39]	53.8	42.0	79.5	59.9
OSTrack [40]	51.5	41.2	72.9	54.9
CAT [18]	45.0	31.4	80.4	56.1
FANet [50]	44.1	34.0	78.7	55.31

Table 1: Comparisons on **RGB-T tracking.**

In this section, we first describe the details of the experimental training and inference process of the proposed PSMTrack. Then, we compared PSMTrack with other state-of-the-art methods on multiple benchmark datasets. Our tracker is evaluated for RGB-T tracking on LasHeR [19] and RGBT234 [17], for RGB-D tracking on DepthTrack [37] and VOT22RGBD [14], and for RGB-E tracking on VisEvent [32]. Finally, we conduct ablation studies.

4.1 Implementation Details

Training. Our model can be used for various modality combinations. We use LasHeR [19] for RGB-T tracking, the training set of DepthTrack [37] for RGB-D tracking, and VisEvent [32] for RGB-E tracking. We train on a single NVIDIA RTX 4090 GPU. In the initial stage, we initialize the backbone with pretrained weights from the RGB tracking foundation model [40]. We train for 50, 30, and

70 epochs on RGB-T, RGB-D, and RGB-E, respectively, with 60,000 pairs of samples in each epoch. We use the AdamW [24] optimizer with a weight decay of 10^{-4} and a learning rate of 10^{-5}.

Inference. During inference, we maintain the same settings as in training. The tracking speed tested on the NVIDIA 4090 GPU is approximately 43.6 frames per second (FPS).

4.2 Comparison with State-of-the-arts

LasHeR. LasHeR [19] is a large-scale RGB-T dataset, containing 1,224 paired RGB and TIR video sequences, with a total of 730,000 image pairs, including 32 target categories and 19 challenge attributes. We adopt Precision Rate (PR) and Success Rate (SR) as evaluation metrics. As shown in Tab. 1, our PSMTrack achieves 76.5% PR and 61.2% SR.

RGBT234. RGBT234 [17] includes 234 pairs of RGB-T video sequences with corresponding ground truths. The evaluation metrics are Maximum Precision Rate (MPR) and Maximum Success Rate (MSR). As shown in the table, our PSMTrack achieves a new state-of-the-art, reaching 91.2% MPR and 67.3% MSR. Meanwhile, the video sequences are annotated with 12 attributes. As shown in Fig. 5, our model can still achieve better performance than other trackers in challenging environments.

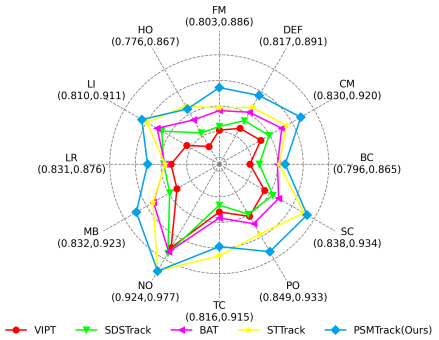

Fig. 5: MPR scores of different attributes on RGBT234 [17].

DepthTrack. DepthTrack [37] is a large-scale RGB-D dataset that covers many challenging real-world scenarios. We use recall (Re), precision (Pr) and F-score as evaluation metrics. As shown in Tab. 2, our PSMTrack attains state-of-the-art performance across all three metrics, which are 64.2%, 64.3%, and 64.2% respectively.

VOT-RGBD2022. The VOT-RGBD2022 [14] dataset is used for RGB-D tracking evaluation and contains 127 sequences. The evaluation metrics include Accuracy, Robustness, and Expected Average Overlap (EAO). The results are

Method	DepthTrack [37]			VOT-RGBD22 [14]		
	Re ↑	Pr ↑	F-score ↑	EAO ↑	Accuracy ↑	Robustness ↑
PSMTrack(Ours)	**64.2**	**64.3**	**64.2**	**77.8**	**82.9**	93.5
STTrack [12]	63.4	63.2	63.3	77.6	82.5	**93.7**
SDSTrack [11]	60.9	61.9	61.4	72.8	81.2	88.3
OneTracker [10]	60.4	60.7	60.9	72.7	81.9	87.2
ViPT [46]	59.6	59.2	59.4	72.1	81.5	87.1
ProTrack [39]	57.3	58.3	57.8	65.1	80.1	80.2
SPT [49]	53.8	52.7	57.8	65.1	79.8	85.1
OSTrack [40]	52.2	53.6	52.9	67.6	80.3	83.3
DET [37]	50.6	56.0	53.2	65.7	76.0	84.5

Table 2: Comparisons on **RGB-D tracking**.

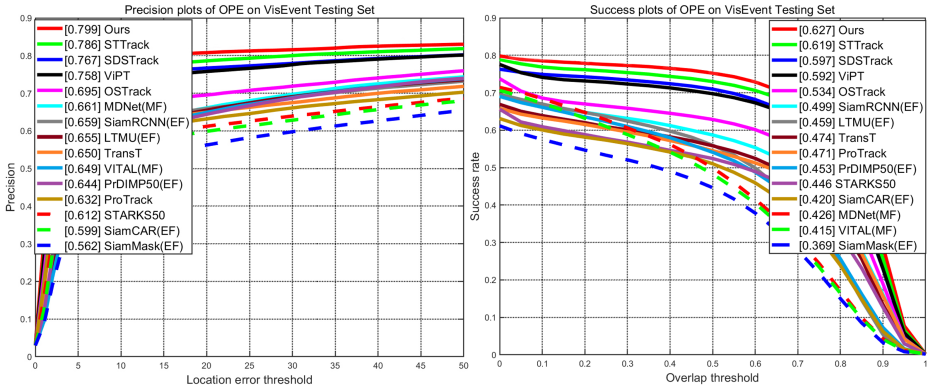

Fig. 6: Overall performance on the VisEvent [32] test set.

shown in Tab. 2. The PSMTrack we proposed achieves an EAO of 77.8%, an Accuracy of 82.9%, and a Robustness of 93.5%, demonstrating its stability and effectiveness under challenging environments.

VisEvent. VisEvent [32] is a dataset currently used for RGB-E tracking, containing 500 pairs of training sequences and 320 pairs of testing sequences. As shown in Tab. 3, our PSMTrack achieves SOTA performances of 62.7% in AUC and 79.9% in precision, respectively. Fig. 6 shows a comparison of the overall performance with other trackers on the VisEvent test set.

	STARK [36]	LTMU [6]	ProTrack [39]	TransT [3]	OSTrack [40]	ViPT [46]	SDSTrack [11]	OneTrack [10]	STTrack [12]	PSMTrack (Ours)
AUC ↑	44.6	45.9	47.1	47.4	53.4	59.2	59.7	60.8	61.9	**62.7**
Pr ↑	61.2	65.9	63.2	65.0	69.5	75.8	76.7	76.7	78.6	**79.9**

Table 3: Comparisons on **RGB-Event tracking**.

4.3 Ablation Studies

Effectiveness of each component. To verify the effectiveness of the proposed modules, we take the complete PSMTrack as the baseline and conduct ablation studies on LasHeR, DepthTrack, and VisEvent, respectively. The experimental results are shown in Tab. 4. Due to the significant differences in target representation across modalities, HSMF can rebalance features at both local and global scales. When this module is removed and replaced with simple feature concatenation, the performance on all three datasets drops noticeably, demonstrating that direct concatenation alone is insufficient to obtain a robust cross-modal representation. Meanwhile, DBME leverages the differential information between modalities and, combined with bidirectional state-space modeling, can effectively enhance the quality of intra-layer representations. In addition, introducing CLI at the end of the network for cross-layer feature interaction can compensate for the information loss caused by relying solely on the last layer features, thereby further improving the overall tracking performance.

Variants	DBME	CLI	HSMF	LasHeR [19]		DepthTrack [37]			VisEvent [32]	
				PR	SR	Re	Pr	F-score	AUC	Pr
Ours	✓	✓	✓	0.765	0.612	0.642	0.643	0.642	0.627	0.799
1	×	✓	✓	0.752	0.605	0.636	0.637	0.636	0.623	0.789
2	✓	×	✓	0.753	0.607	0.638	0.635	0.636	0.624	0.793
3	✓	✓	×	0.748	0.603	0.630	0.634	0.632	0.616	0.775
4	×	×	✓	0.733	0.596	0.626	0.623	0.624	0.614	0.772

Table 4: Ablation studies on the effect of the components. DBME, CLI, and HSMF denote difference-guided bidirectional state-space Mamba enhancement, cross-layer interaction, and hybrid spiral mamba fusion module.

Impact of CLI module design. To further verify the effectiveness of each component within the CLI structure we proposed, we conducted a systematic ablation experiment on its internal components. Taking the original NC-SSD as the baseline, then replace the linear projection with adaptive convolution (AConv), substituted DwConv with SCConv, and add SCSA after the cross-layer concatenation. The experimental results are shown in Tab. 5, indicate that directly using NC-SSD results in weak cross-layer interaction, highlighting that our structural improvement plays a key role in feature alignment and semantic fusion. After using AConv, the model improves the adaptive adjustment ability of feature distributions across different layers, while the addition of SCConv significantly enhances spatial reconstruction capabilities in complex backgrounds, occlusions and weak texture scenes.After further introducing SCSA, the model is able to better suppress spatial redundancy when interacting with multi-level features, optimize the interaction between multi-level features, and achieve the best results on three datasets. In CLI, SCSA, AConv and SCConv work synergistically to establish a stable and efficient cross-layer interaction mechanism, which is crucial for improving the overall tracking performance.

Method	LasHeR	DepthTrack	VisEvent
NC-SSD	60.5	63.6	62.3
+AConv & SCConv	60.9	63.9	62.5
+SCSA	**61.2**	**64.3**	**62.7**

Table 5: Impact of CLI components.

Impact of Top-k Layer Selection. We further investigate the influence of the number of selected layers k in the CLI module. Specifically, the router dynamically selects the top-k layers from all Transformer layers for cross-layer interaction. As shown in Tab. 6, when k is small, the model cannot fully utilize the complementary information from different feature levels; as k increases, the model can fuse richer multi-level features and improve its adaptability to dynamic scenes, thereby enhancing the performance. However, selecting too many layers introduces redundant information and instead degrades the performance. Considering both performance and redundancy suppression, we adopt $k = 3$ in the final model. Through this dynamic selection, the model can more accurately handle different tracking scenarios.

Layers	LasHeR	DepthTrack	VisEvent
1	60.7	63.8	62.4
2	60.9	64.0	62.5
3	**61.2**	**64.3**	**62.7**
4	60.8	64.1	62.5

Table 6: Impact of top-k layer selection in the CLI module.

Visualization results. To intuitively demonstrate the effectiveness of our method, we performed visualizations on multiple datasets. As shown in Fig. 7, we compare our tracker with two other multimodal trackers. The results show that our PSMTrack can efficiently perform multimodal fusion and still achieve more stable and accurate tracking than other methods when dealing with complex scenarios.

5 Conclusion

We propose a new progressive multimodal tracking framework named PSMTrack. By leveraging local Mamba and global spiral Mamba, the model achieves sufficient multimodal feature fusion at the spatial level. In addition, we propose a DBME module, which progressively enhances and interacts with features layer by layer through Mamba and bidirectional state-space modeling. We also introduce a CLI module for cross-layer interaction to alleviate the problem of information loss when relying only on the last layer. Extensive experiments demonstrate that the method we proposed achieves state-of-the-art tracking performance on three multimodal tracking benchmarks.

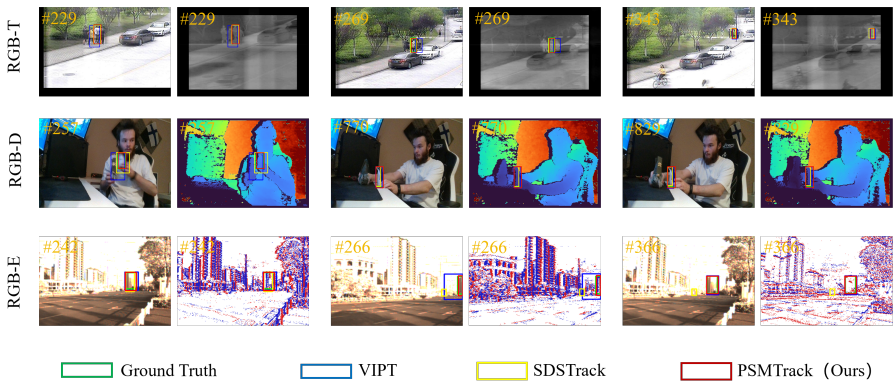

Fig. 7: Visualization results of RGB-T, RGB-D and RGB-E.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant No: 62473205, 62406149, 62363015, Natural Science Foundation of Jiangsu Province No. BK20241893, and sponsored by Yong Academic Leader of Qing Lan Project in Jiangsu Province.

References

1. Cao, B., Guo, J., Zhu, P., Hu, Q.: Bi-directional adapter for multimodal tracking. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 927–935 (2024)
2. Chen, X., Peng, H., Wang, D., Lu, H., Hu, H.: Seqtrack: Sequence to sequence learning for visual object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14572–14581 (2023)
3. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8126–8135 (2021)
4. Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3491–3502 (2022)
5. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13608–13618 (2022)
6. Dai, K., Zhang, Y., Wang, D., Li, J., Lu, H., Yang, X.: High-performance long-term tracking with meta-updater. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6298–6307 (2020)
7. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)

8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
10. Hong, L., Yan, S., Zhang, R., Li, W., Zhou, X., Guo, P., Jiang, K., Chen, Y., Li, J., Chen, Z., et al.: Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19079–19091 (2024)
11. Hou, X., Xing, J., Qian, Y., Guo, Y., Xin, S., Chen, J., Tang, K., Wang, M., Jiang, Z., Liu, L., et al.: Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26551–26561 (2024)
12. Hu, X., Tai, Y., Zhao, X., Zhao, C., Zhang, Z., Li, J., Zhong, B., Yang, J.: Exploiting multimodal spatial-temporal patterns for video object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3581–3589 (2025)
13. Kang, B., Chen, X., Lai, S., Liu, Y., Liu, Y., Wang, D.: Exploring enhanced contextual information for video-level object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4194–4202 (2025)
14. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Chang, H.J., Danelljan, M., Zajc, L.Č., Lukežič, A., et al.: The tenth visual object tracking vot2022 challenge results. In: European Conference on Computer Vision. pp. 431–460. Springer (2022)
15. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
16. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(12), 5743–5756 (2016)
17. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. Pattern Recognition **96**, 106977 (2019)
18. Li, C., Liu, L., Lu, A., Ji, Q., Tang, J.: Challenge-aware rgbt tracking. In: European conference on computer vision. pp. 222–237. Springer (2020)
19. Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., Sun, D.: Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing **31**, 392–404 (2021)
20. Li, J., Wen, Y., He, L.: Scconv: Spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6153–6162 (2023)
21. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European conference on computer vision. pp. 237–255. Springer (2024)
22. Li, X., Zhong, B., Liang, Q., Li, G., Mo, Z., Song, S.: Mambalct: Boosting tracking via long-term context state space model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4986–4994 (2025)
23. Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking. Advances in Neural Information Processing Systems **35**, 16743–16754 (2022)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Luo, Y., Guo, X., Dong, M., Yu, J.: Rgb-t tracking based on mixed attention. arXiv preprint arXiv:2304.04264 (2023)

26. Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. Advances in neural information processing systems **35**, 2846–2861 (2022)
27. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
28. Shi, Y., Li, M., Dong, M., Xu, C.: Vssd: Vision mamba with non-causal state space duality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10819–10829 (2025)
29. Si, Y., Xu, H., Zhu, X., Zhang, W., Dong, Y., Chen, Y., Li, H.: Scsa: Exploring the synergistic effects between spatial and channel attention. Neurocomputing **634**, 129866 (2025)
30. Sun, C., Zhang, J., Wang, Y., Ge, H., Xia, Q., Yin, B., Yang, X.: Exploring historical information for rgbe visual tracking with mamba. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6500–6509 (2025)
31. Tuscher, M., Hörz, J., Driess, D., Toussaint, M.: Deep 6-dof tracking of unknown objects for reactive grasping. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 14185–14191. IEEE (2021)
32. Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., Wu, F.: Visevent: Reliable object tracking via collaboration of frame and event flows. IEEE Transactions on Cybernetics **54**(3), 1997–2010 (2023)
33. Wang, X., Wang, S., Tang, C., Zhu, L., Jiang, B., Tian, Y., Tang, J.: Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19248–19257 (2024)
34. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9697–9706 (2023)
35. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19300–19309 (2024)
36. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10448–10457 (2021)
37. Yan, S., Yang, J., Käpylä, J., Zheng, F., Leonardis, A., Kämäräinen, J.K.: Depthtrack: Unveiling the power of rgbd tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10725–10733 (2021)
38. Yang, J., Li, Z., Yan, S., Zheng, F., Leonardis, A., Kämäräinen, J.K., Shao, L.: Rgbd object tracking: An in-depth review. arXiv preprint arXiv:2203.14134 (2022)
39. Yang, J., Li, Z., Zheng, F., Leonardis, A., Song, J.: Prompting for multi-modal tracking. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3492–3500 (2022)
40. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European conference on computer vision. pp. 341–357. Springer (2022)
41. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. Acm computing surveys (CSUR) **38**(4), 13–es (2006)

42. Zhang, P., Wang, D., Lu, H., Yang, X.: Learning adaptive attribute-driven representation for real-time rgb-t tracking. International Journal of Computer Vision **129**(9), 2714–2729 (2021)
43. Zhang, P., Zhao, J., Wang, D., Lu, H., Ruan, X.: Visible-thermal uav tracking: A large-scale benchmark and new baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8886–8895 (2022)
44. Zhang, Z., Fidler, S., Urtasun, R.: Instance-level segmentation for autonomous driving with deep densely connected mrfs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 669–677 (2016)
45. Zheng, A., Wang, Z., Chen, Z., Li, C., Tang, J.: Robust multi-modality person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3529–3537 (2021)
46. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9516–9526 (2023)
47. Zhu, J., Tang, H., Cheng, Z.Q., He, J.Y., Luo, B., Qiu, S., Li, S., Lu, H.: Dcpt: Darkness clue-prompted tracking in nighttime uavs. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 7381–7388. IEEE (2024)
48. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
49. Zhu, X.F., Xu, T., Tang, Z., Wu, Z., Liu, H., Yang, X., Wu, X.J., Kittler, J.: Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3870–3878 (2023)
50. Zhu, Y., Li, C., Tang, J., Luo, B.: Quality-aware feature aggregation network for robust rgbt tracking. IEEE Transactions on Intelligent Vehicles **6**(1), 121–130 (2020)
51. Zhu, Z., Hou, J., Wu, D.O.: Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22045–22055 (2023)